NBA Final Project Report

Group Members: Andrew Vernon, Jackson Milner, Nahome Kassa

The goal of this project is to analyze NBA player data to find how a player’s physical characteristics impact their draft order, success in the league, and which performance statistics tended to be grouped together. This topic was chosen because we are all fans of basketball and we would like to get a statistical insight to the underlying factors of what makes a player valuable in the league and to be able to use these insights to make new predictions about new players.

The dataset that we have selected contains data on all players who have been on an NBA roster from the years 1996-2019 and has 22 columns of data. The data on the players includes physical characteristics such as age, height, and weight. The data also contains the player’s country of birth, college team played for, draft year, and draft round. Player performance statistics such as the average number of points, rebounds (offensive and defensive), and assists. The dataset also includes player efficiency statistics such as shooting efficiency, rebound percentage, assist percentage, and team play percentage.

Some initial observations about the data are as follows:

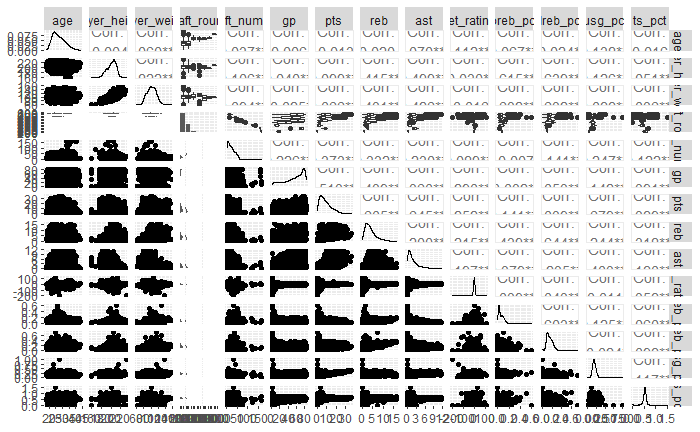
* Contains data on over 11,000 players, so we are comfortable on the quality of data
* Observations contain very specific information, like distinguishing between offensive and defensive rebounds and allocating a column to each of them
* Data on players dates back to 1996, so we can take all of our conclusions from one era and compare it to another and see what’s different and what’s the same

The research questions we have decided to focus on are as follows:

* **How does the player’s college team impact his draft selection?** We have seen that big name basketball schools tend to attract more talented players out of high school. With a conclusion on this research question, we want to be able to determine if where a player went to school has a significant effect on his draft position
* **How much does a player’s draft selection impact his individual game stats?** We want to see if there are any trends with a player’s individual game stats and his draft position. With a conclusion on this research question, we want to be able to predict how good a player’s individual games stats will be based on his draft position.
* **How much have height, age, and weight changed over time?** The sport of basketball has evolved over the years, and recently in the past 5-10 years there has been a greater emphasis on 3-point shooting. Due to the change in the sport, we are seeing more skilled players get selected ahead of taller, stronger, more physically impressive players. With a conclusion on this research question, we want to make additional insights on how valuable a player is to their team.
* **How much does a certain era affect a player’s individual game stats?** As we know, the game of basketball is constantly evolving and the considered “best” way to play basketball is continually in question. We want to know how much a player’s individual game stats are affected by the era that he is playing in.
* **How much effect does draft selection impact how long a player plays in the NBA?** We are interested in whether or not higher drafted players play longer on average than players drafted later or players who went undrafted.

Normality Testing

As an early step in analyzing the data, we were interested in which columns had any correlation with each other as well as the normality of the numerical columns. Using the R function ggpairs, we created a matrix of all the numerical columns plotted against each other as well as the normality of each numerical column.



The only numerical columns that appear to be normally distributed are age, height, and weight. All other numerical columns are either skewed left or skewed right. This makes sense, given that basketball has many positions that are designed for specific roles which certain statistics are associated with. Moreover, the NBA is a star-driven league, so intuitively a small group of players would be have certain stats that are very high compared to the average player, skewing the normality graph in the appropriate direction, depending on the statistic.

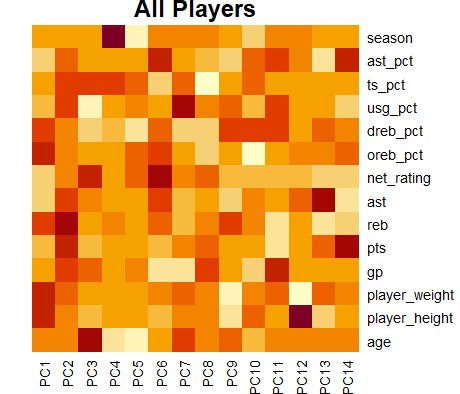
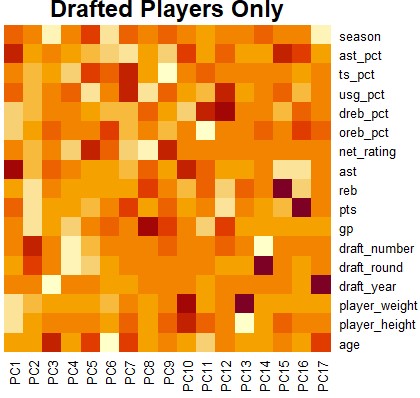
MANOVA

Next, we wanted to prepare the data for the tests that we were planning to do. The first step we took was creating a new column that categorized players based on era played in. To do this, we grouped players into four groups based on a specific year range and assigned them a corresponding numerical value from one to four to represent that category. This preparation step was useful for answering the research questions with regard to a player’s era. We decided on using MANOVA to compare players based on era, because this would show differences in the observations in each statistical category.

The MANOVA test gave us a p-value of 2.2e-16; since the p-value is really small, we reject the null hypothesis. Based on our findings, it is clear to see that the NBA is going younger and slightly shorter. Our observation makes sense because over the last decade, we are seeing an emphasis on three-point shooting as well as emphasis on developing younger NBA players.

PCA

Following this, we wanted to address our research questions pertaining to a player’s draft position and draft round. Because there were a good number of players that were undrafted, we had to remove those players from the dataset to be able to perform our desired methods. After we had done this, we were able to do PCA and include the draft number, draft round, and draft year for players that were drafted. For comparison, we wanted to do PCA again but on all players including the undrafted ones, but not accounting for any of the draft statistics. Below are heatmaps that detail the principal components for each. The lighter color corresponds to negative values, while the darker values correspond to positive values.



It’s important to mention as well the cumulative proportion of variance for each principal component for both of these sets. Because of relevance, we are only showing the first nine principal components.

Cumulative Proportion of Explained Variance

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 |
| Draft only | 0.2439 | 0.4546 | 0.5800 | 0.67021 | 0.74302 | 0.79678 | 0.83926 | 0.88038 | 0.91267 |
| All players | 0.2844 | 0.5125 | 0.6027 | 0.68592 | 0.75058 | 0.80382 | 0.85500 | 0.89652 | 0.9251 |

From the heatmap, we can see the variables that are grouped into the first few principal components for each set of results are very different. For only the drafted players, assists and assist percentage were by far the most important variables in the positive direction, while height and weight were the most important in the negative direction. For the second principal component in the positive direction, draft round as well as draft year were by far the most important variables and in the negative direction, rebounds points, and games played were the most important. For all players, offensive rebound percentage, defensive rebound percentage, total rebounds, height, and weight all had a significant part in the positive direction for the first principal component, while net rating, assist percentage, and assists made up the majority of the negative direction. For the second principal component, individual game statistics assists, rebounds, points, shooting percentage, and team play percentage made up the majority of the positive direction, and there were no significant variables in the negative direction.

The cumulative proportion of explained variance for each principal component was similar for both groups of principal components. For both, the first two principal components made up approximately half of the explained variance. After the first two principal components, there is a drop off in amount of explained variance per principal component. Conclusively, there appears to be two majority groups for both the drafted players only and all players.

By selecting the first 6 principal components, we are able to explain very close to 80% of the total variance. Alternatively, we could select 3 couple more principal components for a total of 9 to explain over 90% of the total variance. PCA is a viable reduction method for this dataset because as with most sports statistics, many of the variables are correlated. This reduction method maximizes the variances explained by the new variables, so it's a useful technique when there is a lot of correlation between the variables.

Clustering

To further explore this dataset, we used clustering to identify similarities and differences between the performances of undrafted and drafted players, and to see if there were any patterns between player draft number and performance. The first thing we did was look at the dataset as a whole to get a general idea of how the players were grouped. We used k-means clustering to identify 5 clusters because there are 5 players on the court for a time at any given time and we thought that this might group the players into the 5 different positions. Looking at the assist vs player height clusters gives us a good idea of how the players were clustered, as the short players with high assists tend to be in the same cluster and tall players with high assists were in the same cluster. Clustering also identified groups of players that were drafted high and performed well, drafted high and performed poorly, drafted late and performed well, and drafted late and performed poorly. There wasn’t much difference between the clustering of the undrafted and drafted players.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Research Questions We Couldn’t Answer

This data analysis has only scratched the surface on the potential of information that could be gathered from this dataset. There were two research questions that we were unable to answer in this research. We couldn’t find if a player’s college had a significant impact on the player’s draft position, because there were two many colleges and we found it impossible to realistically categorize them meaningfully. The other research question we couldn’t answer was how does a player’s draft selection impact how long he plays in the league? We were unable to answer this question because the dataset provided insufficient information on years played.